nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-021-00876-8

Check for updates

# Exome sequencing in obsessive–compulsive disorder reveals a burden of rare damaging coding variants

Mathew Halvorsen[1], Jack Samuels[2], Ying Wang[3], Benjamin D. Greenberg[4], Abby J. Fyer[5], James T. McCracken[6], Daniel A. Geller[7], James A. Knowles[8], Anthony W. Zoghbi[5,9], Tess D. Pottinger[9], Marco A. Grados[2], Mark A. Riddle[2], O. Joseph Bienvenu[2], Paul S. Nestadt[2], Janice Krasnow[2], Fernando S. Goes[2], Brion Maher[10], Gerald Nestadt[2,11] ✉ and David B. Goldstein[9,11] ✉

**Obsessive–compulsive disorder (OCD) affects 1–2% of the population, and, as with other complex neuropsychiatric disorders, it is thought that rare variation contributes to its genetic risk. In this study, we performed exome sequencing in the largest OCD cohort to date (1,313 total cases, consisting of 587 trios, 41 quartets and 644 singletons of affected individuals) and describe contributions to disease risk from rare damaging coding variants. In case–control analyses ($n = 1{,}263/11{,}580$), the most significant single-gene result was observed in *SLITRK5* (odds ratio (OR) = 8.8, 95% confidence interval 3.4–22.5, $P = 2.3 \times 10^{-6}$). Across the exome, there was an excess of loss of function (LoF) variation specifically within genes that are LoF-intolerant (OR = 1.33, $P = 0.01$). In an analysis of trios, we observed an excess of de novo missense predicted damaging variants relative to controls (OR = 1.22, $P = 0.02$), alongside an excess of de novo LoF mutations in LoF-intolerant genes (OR = 2.55, $P = 7.33 \times 10^{-3}$). These data support a contribution of rare coding variants to OCD genetic risk.**

OCD is a neuropsychiatric condition characterized by persistent, intrusive thoughts (obsessions) and repetitive, intentional behaviors (compulsions). The disorder affects approximately 1–2% of the population, with onset in most cases occurring in childhood, adolescence or early adulthood. Evidence from family-based studies supports a genetic contribution to the disorder, with a recent estimate of monozygotic twin-based OCD diagnostic correlation of 0.531 (refs. [1,2]). Genome-wide association studies of common single-nucleotide polymorphisms (SNPs) have not found variants that were associated with OCD at the genome-wide level of statistical significance, likely owing to insufficient sample size, but have reported an SNP-based heritability value of 0.28, consistent with a contribution of these common variants to risk[3–5]. It is plausible that the gap between the twin-based heritability estimate and the SNP-based common variant heritability is due, in part, to rare variation[6].

The contributions to OCD genetic risk from rare coding single-nucleotide variants (SNVs) and insertions–deletions (indels) have been underexplored relative to that of other neuropsychiatric disorders. Recent work has suggested that these variants play a role in the overall genetic architecture. Cappi et al.[7] analyzed de novo SNVs and indels across 184 OCD trios and reported a burden of damaging mutations relative to controls. In addition, rare variant studies of Tourette's syndrome (highly comorbid with OCD) have identified risk genes at genome-wide significance within de novo variant and case–control contexts and similar to Cappi et al.[7] detected a similar burden of damaging de novo mutations (DNMs) in cases relative to controls[8,9].

To assess the burden of rare coding SNVs and indels in OCD, we studied the largest exome-sequenced dataset for this disorder to date. The data included 1,313 total cases from 587 sporadic OCD trios, 41 quartets and 644 additional unrelated OCD cases (Fig. 1). The results of the analysis support a contribution of rare damaging coding variation to OCD risk.

## Results

**Study participants.** All OCD case participants were ascertained as part of the OCD Genetics Association Study, which was described previously[3], or at the Johns Hopkins University OCD clinic. The trios and quartets included in the current study consisted of unaffected parents and, where possible, families with no additionally known affected relatives. See Methods for additional details regarding recruitment, assessment and sequencing. A table of OCD cases and unaffected family members included in this study is provided (Supplementary Table 1).

Controls were largely selected for use from a library of processed samples housed at Columbia University's Institute for Genomic Medicine (IGM). All of IGM's control samples selected were

[1]Department of Genetics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, USA. [2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [3]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [4]Department of Psychiatry and Human Behavior, Brown Medical School, Providence, RI, USA. [5]New York State Psychiatric Institute, College of Physicians and Surgeons at Columbia University, New York, NY, USA. [6]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine at Los Angeles, Los Angeles, CA, USA. [7]Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [8]SUNY Downstate Medical Center College of Medicine, Brooklyn, NY, USA. [9]Institute for Genomic Medicine, Columbia University Medical Center, New York, NY, USA. [10]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA. [11]These authors contributed equally: Gerald Nestadt, David B. Goldstein. ✉e-mail: gnestadt@jhmi.edu; dg2875@cumc.columbia.edu

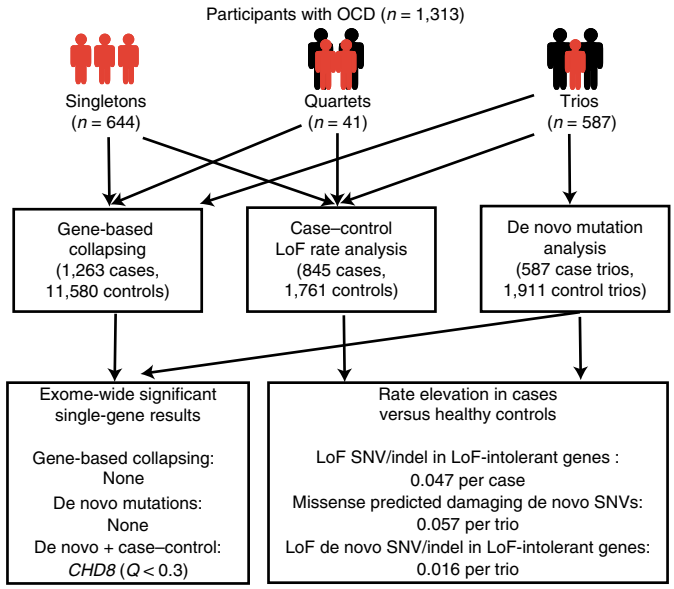


**Fig. 1 | Overview of analysis design.** Sequence data for a total of 1,313 cases were used across a variety of family structures (644 unrelated cases, 41 quartets and 587 trios). Although singleton cases had a variety of family histories, trios and quartets were targeted for a childhood age at onset (<18 years) and a lack of family history. We constructed three separate analysis cohorts from these data: (1) 1,263 unrelated cases and 11,580 unrelated controls negative for an explicit psychiatric phenotype for gene-based collapsing analysis; (2) 845 unrelated European ancestry cases and 1,761 unrelated European ancestry healthy controls for exome-wide comparisons of LoF rate; and (3) 587 OCD case trios and 1,911 healthy control trios for DNM analyses. On the lower left, single-gene findings from the study are highlighted; on the lower right, estimated rate elevations per OCD case are reported across critical subsets of rare damaging coding variation.

approved for such studies and did not carry a specific defined neuropsychiatric phenotype. For comparisons of damaging coding variant burden across the entire exome, only controls with a reported healthy phenotype were used. For analyses focused specifically on trios, we used published data from the Simons Simplex Collection, across 1,911 unaffected siblings of autism spectrum disoder (ASD) probands[10].

**Analysis design.** We constructed three main analysis cohorts to assess different subsets of the contribution of coding SNVs and indels to OCD (Fig. 1). The gene-based collapsing analysis cohort was designed to deliver optimal power in testing for the presence of single genes with an exome-wide significant burden of rare damaging SNVs and indels in OCD cases relative to controls across separate sample ancestry groups. The LoF rate analysis cohort was designed to compare the rate of rare LoF variation between a single-ancestry grouping of OCD cases and controls listed as healthy. The trio cohort was specifically designed to assess a contribution from the subset of rare coding SNVs and indels that are de novo in origin. A breakdown of case–control sample sizes, kits used, coverage and other statistics is provided (Supplementary Table 2), along with a breakdown of broadly defined phenotypes per control analysis cohort (Supplementary Table 3).

**Gene-based collapsing analyses.** We first constructed a series of formal gene-based collapsing analyses of damaging coding variant burden in cases versus controls. This approach defined qualifying variation based on annotation, call quality control (QC) and case–control coverage similarity and then, for each gene tested, for a difference in the proportion of samples that are carriers in cases versus controls using a two-sided Fisher's exact test. Summary statistics were meta-analyzed using a two-sided Cochran–Mantel–Haenszel test across single-group contingency tables. We excluded samples with poor QC or cryptic relatedness, and, after conducting principal component analysis (PCA) on the common variant genotype data, we used Louvain clustering on principal components (PCs) 1–6 to define separate groups of samples whose clustering reflects similar ancestry (Methods). A total of 11 case–control groups defined according to Louvain clustering were constructed, consisting of 1,263 cases and 11,580 controls in total (Supplementary Table 2).

We defined two sets of gene-based collapsing analyses. The first was focused specifically on LoF variants (Fig. 2a and Supplementary Table 4). The second focused on broader damaging coding annotation and includes LoF variants, in-frame indels and missense predicted damaging variants defined via PolyPhen-2 (ref. [11]) HumDiv $\geq 0.957$, referred to hereafter as misD (Fig. 2b and Supplementary Table 5). All variants were very rare in the general population (maximum population allele frequency (AF) < 0.01% across the case–control group and gnomAD subpopulations that were negative for a neurological phenotype). We set the significance threshold based on Bonferroni adjustment as 0.05 / (18,816 genes × 2 sets of tests) $= 1.3\times10^{-6}$. Scatter plots from the first two PCs of each group are provided (Supplementary Fig. 1).

No single protein-coding gene passed the exome-wide significance threshold. The most significant single-gene result observed across these tests was in *SLITRK5* under the damaging coding model (OR = 8.8, 95% confidence interval (CI), 3.4–22.5, $P = 2.3\times10^{-6}$; Fig. 2b). There was no clear clustering of missense variation across case variants in this gene (Supplementary Fig. 3). The use of missense constraint metrics[12–14] to enrich for more deleterious variation is hampered by sample size and does not lead to full partitioning of case–control missense variation (Supplementary Table 7). A table of qualifying *SLITRK5* variants in cases and controls is provided (Supplementary Table 6), alongside alignments underlying case *SLITRK5* variant calls (Supplementary Fig. 2).

**Case–control comparisons of LoF variant rate.** We compared the rate of LoF variants within a single-ancestry set of 845 OCD cases and 1,761 explicitly healthy controls to determine if OCD, like other psychiatric disorders, has an elevated LoF rate within LoF-intolerant genes (Methods). All samples come from three neighboring clusters in the gene-based collapsing cohort that form the core of self-described individuals of European ancestry (Supplementary Figs. 1 and 4). LoF SNV/indel rate was compared across ten deciles of the recently described LOEUF gene-level score for LoF intolerance[15]. Scatter plots of the first eight PCs from an analysis of common variation in this group are provided (Supplementary Fig. 4).

We found evidence that cases carry an elevated LoF rate within the most intolerant LOEUF decile of genes (OR = 1.33, $P = 0.01$; Fig. 3). In contrast, there is no evidence for a difference in case–control LoF rate in any other decile. The elevation in this LOEUF decile is present in a simpler case–control comparison of LoF versus synonymous variant count (OR = 1.26, $P = 0.05$; Supplementary Fig. 5). Based on linear regression of case–control LoF rate in the most intolerant LOEUF decile, there is an estimated excess of 0.047 rare LoF SNVs and indels within this bin per OCD sample relative to controls ($P = 0.01$; Supplementary Fig. 5). This enrichment is more significant in genes that are defined as LoF intolerant using a threshold of pLI > 0.995 as in similar recent work[16] (OR = 1.38, $P = 3.86\times10^{-3}$; Supplementary Fig. 6). Case LoF enrichment relative to controls is particularly pronounced within genes that are in the most intolerant LOEUF decile and are LoF intolerant (pLI > 0.995), specifically within the non-psychiatric subset of ExAC (OR = 1.56, $P = 1.04\times10^{-3}$; Supplementary Fig. 6).

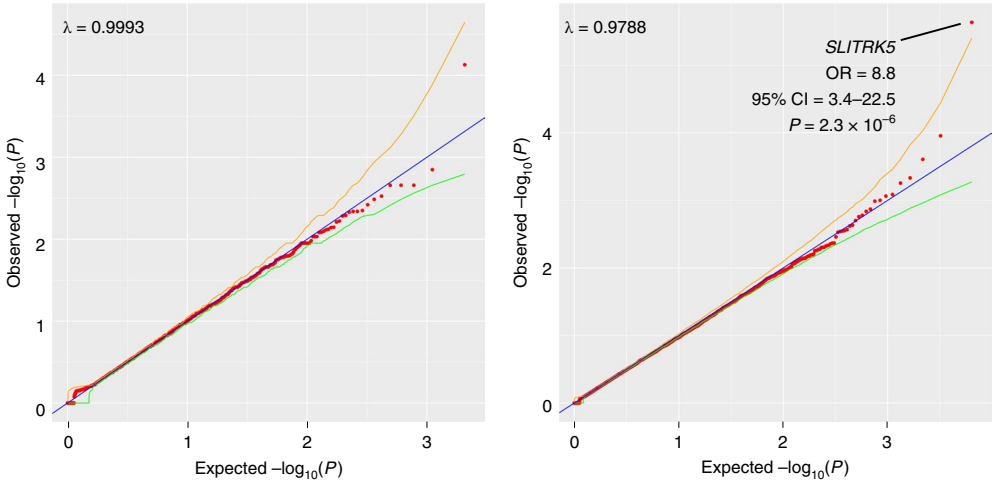


**Fig. 2 | QQ plot from gene-based collapsing meta-analysis across 11 separate case–control clusters that reflect ancestry (total case–control *n* = 1,263/11,580). a**, Gene-based test statistics under an LoF-dominant model. **b**, Test statistics under an LoF and damaging missense dominant model. The summary statistics of *SLITRK5* under this model are the most significant observed in these tests but do not pass the exome-wide significance threshold (Cochran–Mantel–Haenszel chi-squared test (OR = 8.8, 95% CI, 3.4–22.5, two-sided unadjusted $P = 2.3 \times 10^{-6}$)).

**Analysis of de novo SNVs and indels.** The evidence for excess LoF SNVs and indels within LoF-intolerant genes suggests that a study specifically focused on the subset of rare coding variants that likely confer the highest degree of relative risk is worthwhile. In other complex neuropsychiatric disorders, the variants that confer the largest relative risk are often de novo in origin, because these variants have not been subjected to purifying selection. We therefore tested for an elevated rate of damaging (LoF and misD as defined above) DNMs in OCD trio sequence data.

**De novo SNV and indel calling.** We called de novo SNVs and indels across a total of 587 trios and 41 quartets (Supplementary Tables 8 and 9). Candidate calls were required to meet stringent QC thresholds including (1) joint coverage ≥ 10×; (2) alternate AF ≥ 30% in the proband; and (3) alternate AF < 5% in the reads of each parent. Alignments underlying each QC-passing DNM call were visually inspected[17], and calls that failed visual inspection were excluded from the final callset. Coverage statistics suggest that our data were well powered for calling across the exome and that we captured the bulk of DNMs in these loci (Methods and Supplementary Tables 10 and 11). Assessable results from selected SNVs and all frameshift indel calls sent for Sanger sequencing resulted in a validation rate of 80/81 (98.8%) and 22/23 (95.7%), respectively, suggesting that almost all variants remaining that either did not undergo validation or were not assessable are likely real (Methods and Supplementary Tables 12 and 13).

**DNM burden.** We compared the burden of DNMs per annotation in the cohort of 587 sporadic OCD trios relative to 1,911 control trios. Control DNMs are from a previously published callset[10] and underwent the same type of variant annotation as case DNMs. misD and LoF variation are defined as they were in earlier case–control analyses. To control for differences in coverage, we defined jointly covered loci as covered at least 10× in more than 90% of an internal set of 709 controls sequenced on the same kit as control trios (Roche EZCap v2), as well as more than 90% of OCD trio probands. All case–control tests were conducted using the total number of DNMs outside these loci per sample as a covariate. We also constructed additional complementary comparisons of DNM rate to expectation based on a previously described framework[18,19], incorporating cohort size, gene size and sequence content (Methods). DNM calls from all cases and controls alongside their assigned coding annotations are provided in full (Supplementary Table 14).

Across the exome, we note an excess of misD DNMs in OCD cases relative to controls (OR = 1.22, $P = 0.02$; Fig. 4a). In comparison, there was no difference in the burden of presumably neutral synonymous and missense non-predicted damaging (misND) mutations. The rate increase of misD SNVs relative to controls was 0.057 ($P = 0.02$; Supplementary Fig. 7). These same patterns are observed relative to expected mutation rate based on sequence context (Supplementary Fig. 8).

As observed in case–control comparisons of LoF SNV and indel burden described earlier, the case burden of LoF DNMs relative to expectation was concentrated specifically within the most intolerant genes as defined by the LOEUF decile (OR = 2.55, $P = 7.33 \times 10^{-3}$; Fig. 4b). This burden represents a rate excess in cases relative to controls of 0.016 ($P = 4.86 \times 10^{-3}$; Supplementary Fig. 7). The rate excess was also present relative to expectation based on mutation rate (Supplementary Fig. 8). The observed rate was not significantly different from the expected rate outside of LoF-intolerant genes for both LoF SNVs and LoF indels (Supplementary Fig. 9).

We note a distinct excess of LoF/misD DNMs in genes that are relevant to OCD (Supplementary Fig. 10). In a set of broad neurodevelopmental genes ($n = 187$ (refs. [16,20])), a rate excess similar to that of LoF variants in LoF-intolerant genes was present (observed/expected = 10/4.93, rate ratio = 2.03, Poisson $P = 0.03$). There is a strong enrichment of LoF/misD DNMs within risk genes for Tourette's syndrome ($n = 6$ (ref. [8])), known to be highly comorbid with OCD (observed/expected = 4/0.37, rate ratio = 10.7, Poisson $P = 6.04 \times 10^{-4}$).

**Burden of misD de novo SNVs in intolerant coding regions.** There was some evidence that cases are more likely to carry misD DNMs within bins of high missense constraint, as defined by the Missense badness, PolyPhen-2 and Constraint (MPC) score. Case misD DNM burden relative to controls is not concentrated in the most intolerant LOEUF decile (Supplementary Fig. 11), suggesting that a missense metric such as MPC that incorporates regional constraint is necessary for analyses of these variants. There was not a significant excess of misD DNMs within the MPC > 2 bin alone (OR = 1.71, $P = 0.13$). We found that the combined burden of these DNMs and LoF DNMs in LoF-intolerant genes in cases versus controls was more significant than the burden for either annotation alone (OR = 2.12, $P = 2.75 \times 10^{-3}$; Supplementary Fig. 12).

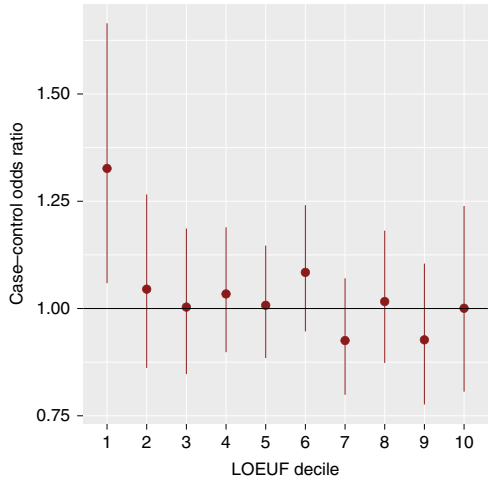


**Fig. 3 | Enrichment of LoF variation across LOEUF deciles in 845 cases relative to 1,761 healthy controls (dots), with 95% CIs provided (bars).** Enrichment was calculated using a two-sided logistic regression model, with phenotype regressed on the number of LoF variants per individual alongside covariates described in the main text. Results suggest that OCD cases are more likely to carry LoF variation, specifically within the first LOEUF decile (that is, the most LoF-intolerant set of genes in the genome) compared to healthy controls (OR = 1.33, unadjusted $P = 0.01$).

**Sex bias in OCD cases with the most damaging de novo SNVs and indels.** There was evidence for a relationship between carrier status of the most deleterious DNMs (LoF in a gene with LOEUF < 10% or misD with MPC > 2) and sex. We tested for an association between carrier status of one of these DNMs and four separate binary phenotypes (male sex, tics, skin-picking and trichotillomania) and found that only male sex was an effective predictor of carrier status (72.7% versus 47.3%; OR = 3.19, 95% CI, 1.5–7.7, $P = 0.006$; Supplementary Fig. 13). An independent analysis of LoF variant burden in non-trio male cases versus non-trio female cases indicated a similar male burden specific to genes with LOEUF < 10% (OR = 1.75, $P = 0.009$; Supplementary Fig. 14).

**Gene-based studies of DNM rate.** We put together two sets of gene-based tests of damaging DNM rate relative to expectation. The first set of tests focused on DNMs with LoF or misD annotation, whereas the second focused on LoF DNMs specifically. We set the Bonferroni-corrected $P$ value threshold for significance as 0.05 / (18,852 genes tested × 2 sets of tests) = $1.3 \times 10^{-6}$. We added power to these tests by integrating a DNM callset for 184 separate OCD trios from Cappi et al.[7], bringing the total cohort size to 771.

In formal DNM rate tests, we failed to detect an individual gene that reached an exome-wide level of significance (Supplementary Table 15). Among the top-ranking results, we detected three LoF/misD DNMs in *CHD8* across the combined cohort, two previously reported and described in Cappi et al.[7] and one detected within our 587 trios (Poisson $P = 7.16 \times 10^{-5}$). The new mutation identified in our cohort was a nonsense SNV and was Sanger validated (Supplementary Fig. 15). *CHD8* is a neurodevelopmental gene whose dysfunction has already been implicated in autism[10] and developmental disorders[21].

**Joint analysis of DNM and case–control data with extTADA.** We used extTADA[22] to construct a Bayesian analysis focused on LoF and misD DNMs that is bolstered by LoF variant counts from independent OCD case–control data. The case–control LoF variant count data were derived from the previously described Caucasian cohort used for LoF rate comparisons that were shown to carry an excess of LoF variation in LoF-intolerant genes. We removed trio cases from the set of samples used for the previous LoF rate analysis to obtain a set of 476 cases and 1,761 controls that could be jointly analyzed with the 771 trios using extTADA (Methods). Within the results, there was only one gene with status as a 'probable risk gene' (here defined as significant at a false discovery rate (FDR) of 0.3—that is, $Q < 0.3$): *CHD8* ($Q = 0.22$; Fig. 5 and Supplementary Table 15).

**Notable DNM calls.** We formed a summary table of DNM calls that were highlighted in the analyses described (Supplementary Table 16). These include DNMs that are particularly deleterious based on LoF or misD annotation described earlier, being a part of gene-based signal involving a recurrent LoF/misD hit, or that overlap with highlighted neurodevelopmental disorder or Tourette's syndrome gene sets. It also includes a de novo nonsense SNV in the *HDAC4* gene that was detected in two siblings within a quartet and was Sanger validated (Supplementary Fig. 16). The failure to detect this variant in either parent suggests that it is parental mosaic in origin, and DNMs with this annotation (LoF in an LoF intolerant gene) have already been noted as elevated in OCD trios.

## Discussion

An excess of rare damaging coding variation has been detected across multiple studies focused on different psychiatric cohorts[8–10]. This study is, to our knowledge, the most comprehensive catalogu-

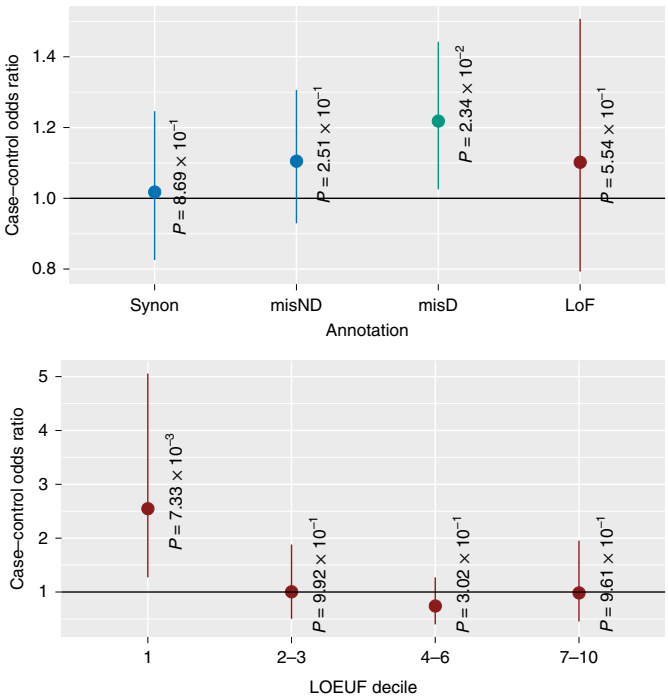


**Fig. 4 | Enrichment of coding DNMs in 587 OCD trios relative to 1,911 healthy controls (dots), with 95% CIs for the estimates provided (bars) and two-sided unadjusted *P* values (right of dots).** Estimates for enrichment were calculated from a logistic regression model of case–control outcome versus DNM counts, with out-of-joint loci DNM counts as covariates. **a**, Case–control enrichment of DNMs partitioned by variant annotation. The only annotation whose burden is most significantly associated with OCD case status is that of missense variants predicted to be damaging (misD). **b**, Case–control enrichment of LoF DNMs across the exome, partitioned by genic depletion of LoF variation in the general population. As with the case–control comparison, only genes that are LoF-intolerant (LOEUF < 10%) carry an excess of LoF DNMs in OCD relative to controls.

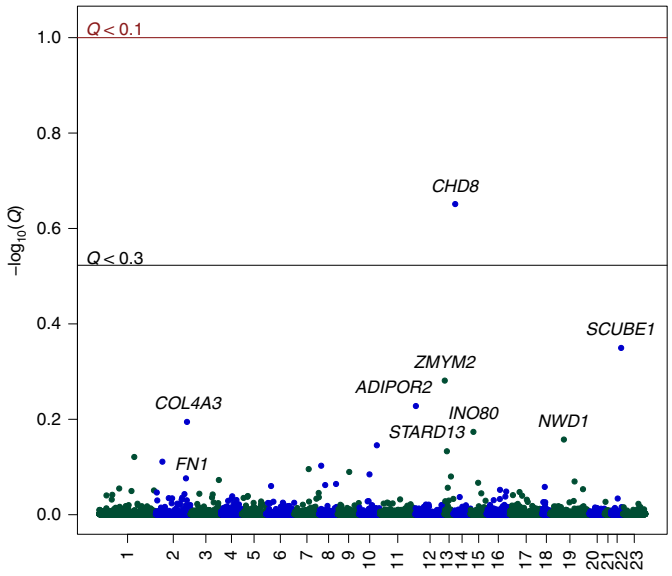


**Fig. 5 | Results of gene-based tests combining DNM data and independent case–control data.** Here, information from LoF and misD de novo SNVs and indels from a total of 771 OCD trios (587 from this study and 184 published by Cappi et al.[7]) is combined with LoF variant counts from an independent group of 476 cases and 1,761 healthy controls, formed via the same clustering protocol used to obtain the LoF rate group but this time including only singleton cases (Methods). All gene symbols listed are those with at least two LoF or misD DNMs across trios and carry a one-sided Poisson test $P < 0.05$ for observed versus expected mutation rate. No gene can be classified as a high-confidence risk gene ($Q < 0.1$) in these results, and the only gene that can be classified as a probable risk gene ($Q < 0.3$) is *CHD8* ($Q = 0.22$; hit with two LoF DNMs, one misD DNM and zero singleton case or control LoF variants).

ing of contributions to OCD risk from rare damaging coding SNVs and indels thus far. Its findings suggest that, like the genetic architecture of other neuropsychiatric disorders[8,16,23,24], OCD involves contributions to overall risk from these variants.

Although the result does not pass the exome-wide significance threshold, the gene *SLITRK5* carries the most significant excess of case rare damaging coding variants relative to controls in this study. *SLITRK5* is a member of the SLITRK gene family, which influences excitatory and inhibitory synapse formation. The protein products of these genes accomplish this by interacting with LAR-RPTPs (leukocyte common antigen-related receptor protein tyrosine phosphatases) through the LRR1 domain[25–29]. *Slitrk5*-knockout mice have been described as having increased 'OCD-like' behaviors, including elevated anxiety and excessive grooming[30]. In human samples, a burden of *SLITRK5* coding variants that influence synapse formation in vitro has previously been described in OCD cases relative to controls[31]. We note that, in accessing results from the SCHEMA study—an exome sequencing study consisting of around 25,000 schizophrenia cases and 100,000 controls—cases carry an excess of LoF variation in *SLITRK5* (OR = 6.69, five carrier cases versus three carrier controls, $P = 9.17 \times 10^{-4}$)[32]. In ASD trios, a total of four de novo missense mutations have been observed in *SLITRK5*, three of which appear to cluster in a 100-nucleotide region[10,33]. Although none of the *SLITRK5* qualifying variation in the OCD case cohort was classified as de novo in origin, full parental sequence data were available for only around half of these variants, and it is possible that some subset of the case signal from singletons are, in fact, de novo in origin.

The gene *CHD8*, the only probable risk gene detected from analyses centered on DNMs, has underlying biology and previous genetic findings consistent with that of an OCD risk gene. LoF mutations in *CHD8* have been implicated in autism and developmental disorders[10,21]. The gene is a transcriptional regulator of neural development and has already been shown to control dosage of several other genes that have been reported as risk genes for neuropsychiatric disorders[34]. Continued observation of damaging DNMs in additional OCD probands will be critical to the formal implication of *CHD8* dysfunction with OCD.

Among the results of these analyses, an unexpected observation was that male OCD trio probands carried a higher load of damaging de novo coding SNVs and indels than female OCD trio probands. This finding runs counter to that of ASD, where current evidence suggests that female ASD probands have a higher burden of damaging DNMs than male ASD probands[16]. The prevailing theory for this sex bias in the burden of damaging DNMs in ASD is the presence of a generalized female protective effect with regard to the ASD genetic risk[16]. However, given that ASD diagnoses are predominantly male and that OCD diagnoses are much closer to a 50:50 sex ratio, a similar effect in OCD seems less likely. Assuming that this observation is not a chance occurrence, several possible explanations exist. One is that these DNMs lead to what are essentially single-gene neurodevelopmental disorders (for example, *CHD8* LoF) and that these have male bias for a diagnosis, particularly one presenting with OCD. Another set of explanations center on selection of cases for this study. It is possible that the sex imbalance of DNM burden is related to our focus on childhood-onset OCD cases or is influenced by the exclusion of cases with significant comorbidities, such as ASD. OCD is thought to carry an underappreciated level of heterogeneity[35], and results here might highlight subtypes that feature greater sex imbalance and rare variant contribution to risk. Finally, it is possible that the imbalance reflects actual biological differences in OCD etiology between male and female cases. Other OCD genetic studies have described intriguing differences in burden between male and female cases, but, like this study, they were were underpowered to produce a high-confidence result[36,37].

The analyses described have enabled us to empirically estimate the rate of risk variation based on the rate increase of these rare coding variants relative to healthy controls. In patients with OCD overall, we estimate that LoF SNVs and indels in LoF-intolerant genes that contribute to OCD risk occur at a rate of 0.047 per case. In sporadic OCD cases, we estimate that risk-conferring misD DNMs occur at a rate of 0.057 per trio and that risk-conferring LoF DNMs in LoF-intolerant genes occur at a rate of 0.016 per trio. Higher-confidence estimates of these rates, as well as implication of additional risk genes, will involve sequencing (1) more OCD case samples and (2) more sporadic OCD trios for gene-based collapsing and DNM analyses, respectively, as well as joint analyses combining both data types.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00876-8.

Received: 24 February 2020; Accepted: 18 May 2021;
Published online: 28 June 2021

## References

1. Nestadt, G., Grados, M. & Samuels, J. F. Genetics of obsessive–compulsive disorder. *Psychiatr. Clin. North Am.* **33**, 141–158 (2010).
2. Polderman, T. J. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
3. Stewart, S. E. et al. Genome-wide association study of obsessive–compulsive disorder. *Mol. Psychiatry* **18**, 788–798 (2013).

4. Mattheisen, M. et al. Genome-wide association study in obsessive–compulsive disorder: results from the OCGAS. *Mol. Psychiatry* **20**, 337–344 (2015).
5. International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) & OCD Collaborative Genetics Association Studies (OCGAS). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol. Psychiatry* **23**, 1181–1188 (2018).
6. Kendall, K.M. et al. Association of rare copy number variants with risk of depression. *JAMA Psychiatry* **76**, 818–825 (2019).
7. Cappi, C. et al. De novo damaging DNA coding mutations are associated with obsessive–compulsive disorder and overlap with Tourette's disorder and autism. *Biol. Psychiatry* **87**, 1035–1044 (2020).
8. Wang, S. et al. De novo sequence and copy number variants are strongly associated with Tourette disorder and implicate cell polarity in pathogenesis. *Cell Rep.* **24**, 3441–3454 (2018).
9. Willsey, A. J. et al. De novo coding variants are strongly associated with Tourette disorder. *Neuron* **94**, 486–499 (2017).
10. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
11. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
12. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353 (2017).
13. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
14. Traynelis, J. et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* **27**, 1715–1729 (2017).
15. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
16. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
17. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
18. Epi, K. C. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
19. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
20. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
21. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
22. Nguyen, H. T. et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).
23. Howrigan, D. P. et al. Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat. Neurosci.* **23**, 185–193 (2020).
24. Rees, E. et al. De novo mutations identified by exome sequencing implicate rare missense variants in *SLC6A1* in schizophrenia. *Nat. Neurosci.* **23**, 179–184 (2020).
25. Takahashi, H. et al. Selective control of inhibitory synapse development by Slitrk3-PTPδ trans-synaptic interaction. *Nat. Neurosci.* **15**, 389–398 (2012).
26. Um, J. W. & Ko, J. LAR-RPTPs: synaptic adhesion molecules that shape synapse development. *Trends Cell Biol.* **23**, 465–475 (2013).
27. Um, J. W. et al. Structural basis for LAR-RPTP/Slitrk complex-mediated synaptic adhesion. *Nat. Commun.* **5**, 5423 (2014).
28. Yim, Y. S. et al. Slitrks control excitatory and inhibitory synapse formation with LAR receptor protein tyrosine phosphatases. *Proc. Natl Acad. Sci. USA* **110**, 4057–4062 (2013).
29. Han, K. A., Jeon, S., Um, J. W. & Ko, J. Emergent synapse organizers: LAR-RPTPs and their companions. *Int. Rev. Cell Mol. Biol.* **324**, 39–65 (2016).
30. Shmelkov, S. V. et al. Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive–compulsive-like behaviors in mice. *Nat. Med.* **16**, 598–602 (2010).
31. Song, M. et al. Rare synaptogenesis-impairing mutations in *SLITRK5* are associated with obsessive compulsive disorder. *PLoS ONE* **12**, e0169994 (2017).
32. Singh, T. et al. Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.18.20192815 (2020).
33. Guo, H. et al. Genome sequencing identifies multiple deleterious variants in autism patients with more severe phenotypes. *Genet. Med.* **21**, 1611–1620 (2019).
34. Sugathan, A. et al. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc. Natl Acad. Sci. USA* **111**, E4468–E4477 (2014).
35. Nestadt, G. et al. Obsessive–compulsive disorder: subclassification based on co-morbidity. *Psychol. Med.* **39**, 1491–1501 (2009).
36. Wang, Y. et al. Gender differences in genetic linkage and association on 11p15 in obsessive–compulsive disorder families. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **150B**, 33–40 (2009).
37. Khramtsova, E. A. et al. Sex differences in the genetic architecture of obsessive–compulsive disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **180**, 351–364 (2019).

## Methods

**Study participants and procedure of whole-exome sequencing.** Written informed consent (or assent, for children) to study procedures was obtained for all participants. The study was reviewed and approved by the institutional review boards at the five sites where participants were recruited: Butler Hospital, Columbia University, Johns Hopkins University, Massachusetts General Hospital and the University of California at Los Angeles. Participants received $100 for completing a diagnostic interview and provision of a blood sample. Unaffected relatives of probands received $35 for provision of a blood sample. Written informed consent was obtained from the participants. The study was in compliance with all ethical regulations.

All case samples were ascertained as part of the OCD Genetics Association Study or at the Johns Hopkins University OCD clinic. The trios and quartets included in the study consisted of unaffected parents and, where possible, families with no additionally known affected relatives. Full trio and quartet families were screened for evidence of OCD[38]. All affected cases were examined by a PhD research psychologist, using an adapted version of the OCD section of the Schedule for Affective Disorders and Schizophrenia[39] for assessing *Diagnostic and Statistical Manual of Mental Disorders-IV* (DSM-IV) OCD and the Schedule Clinical Interview for DSM-IV[40] for assessing major axis I diagnoses other than OCD. Where possible, an informant also was interviewed. All cases were reviewed independently by two research diagnosticians to establish diagnostic concordance.

OCD samples underwent exome sequencing at Columbia University's IGM with the NimbleGen SeqCap EZ exome enrichment kit version 3.0 (Roche) or the IDT Exome Research Panel kit (Integrated DNA Technologies) using Illumina GAIIx or HiSeq 2000 or HiSeq 2500 sequencers and following standard protocols. An additional small number of OCD case samples underwent whole-genome sequencing at the same site (see Supplementary Table 1 for a catalog of the full cohort).

Once high-throughput sequence data were generated, they were analyzed using the IGM's standard DNA sequence data alignment and variant calling pipeline. All data were processed on a pipeline consisting of primary alignment and duplicate marking using the Dynamic Read Analysis for Genomics (DRAGEN) platform followed by variant calling according to best practices outlined in the Genome Analysis Tool Kit (v3.6) as described previously[41]. Reads were aligned to the human reference genome build 37 with decoy sequences included (hs37d5).

**Initial selection of samples for case–control analyses.** All individual case and control samples that were selected for inclusion in case–control analyses were required to meet stringent QC thresholds based on sequencing metrics. A sample was selected for subsequent analyses if it met the following thresholds: consensus coding sequence (CCDS) mean coverage at least 20×; at least 82% of CCDS bases covered at least 10×; percent read duplication < 30%; at least 70% of reads that align to the reference genome; less than 8% PCR contamination estimated; samples either whole-genome sequenced or exome sequenced on a Roche NimbleGen EZCap version 3, IDT Exome Research Panel version 1, Agilent SureSelect Human All Exon V5 or Illumina TruSeq 65MB kit.

All selected samples that met these described QC thresholds were required to show no evidence of cryptic relatedness with other reportedly unrelated samples across the full data, regardless of ancestry. We used the kinship function of the KING version 1.4 statistical genetics software package on 29,961 common variants found in exome data (targeted by the original curators to the NimbleGen EZcap v3 kit) to determine pairwise relatedness patterns among all qualifying samples with genotype call rate ≥ 0.99 (filter applied using PLINK v1.90_3.38 (ref. [42])) and prune members of cryptic relatedness pairs (classified as relatedness coefficient > 0.0884) in a manner that maximizes the number of cases in the final set of samples. We tried this same kinship analysis on a subset of the variants that have genotype call rate ≥ 0.98 in all input samples and found the kinship-pruned set of samples to be exactly the same.

The first six PCs (generated across 12,840 variants using FLASHPCA v2.0 (ref. [43])) were used to perform Louvain clustering as described previously[44] to obtain ancestry-based clusters. Using these methods, we identified a total of 11 clusters of samples that are usable for gene-based collapsing analyses (Supplementary Fig. 1 and Supplementary Table 2). Gene-based collapsing was performed per cluster to test for differences between an individual diagnosed with OCD with at least one qualifying variant compared to controls. An exact two-sided Cochran–Mantel–Haenszel test was subsequently used to test for associations across clusters[44].

**Gene-based collapsing analyses.** For each analysis-ready case–control group, we identified coding variants deemed suitable for analysis based on calling in coverage-balanced loci and being of high quality. For each group, we first identified the subset of CCDS release 20 coding and splice donor–acceptor bases that meet all of the following criteria: (1) at least one case sample and at least one control sample with coverage at the site ≥ 10× and (2) difference in percent of cases covered at least 10× versus percent of controls covered at least 10× no greater than 0.07. In each group, we extracted from corresponding coverage-harmonized bases all coding variant genotypes that met the following criteria: ATAV filter classification of PASS, LIKELY or INTERMEDIATE (VQSR tranche < 99.9% for SNVs and VQSR tranche < 99% for indels); coverage ≥ 10×; indel FS < 200; snv FS < 60; SNV strand OR < 3; indel strand OR < 10; GQ ≥ 20; MQ ≥ 40; QD ≥ 2; QUAL ≥ 20; RPRS > −4; case–control leave-one-out MAF < 0.0001; gnomAD genome MAF < 0.0001 in non-neuro global, AFR, AMR, ASJ, EAS, FIN and NFE; gnomAD genome snv RF > 0.4; gnomAD genome indel RF > 0.4; gnomAD exome MAF < 0.0001 in non-neuro global, AFR, AMR, ASJ, EAS, SAS, FIN and NFE; gnomAD exome snv RF > 0.1; gnomAD exome indel RF > 0.2; het percent alt read from 30–80%; clinEff v1.0c effect of HIGH, MODERATE or LOW; clinEff v1.0c effect in exon_loss_variant, frameshift_variant, rare_amino_acid_variant, stop_gained, start_lost, stop_lost, splice_acceptor_variant, splice_donor_variant, gene_fusion, bidirectional_gene_fusion, 3_prime_UTR_truncation+exon_loss_variant, 5_prime_UTR_truncation+exon_loss_variant, coding_sequence_variant, disruptive_inframe_deletion, disruptive_inframe_insertion, conservative_inframe_deletion, conservative_inframe_insertion, missense_variant + splice_region_variant, missense_variant, splice_region_variant, 5_prime_UTR_premature_start_codon_gain_variant, initiator_codon_variant, initiator_codon_variant + non_canonical_start_codon, splice_region_variant+synonymous_variant, splice_region_variant, start_retained, stop_retained_variant and synonymous_variant. Using the set of variant calls that met the criteria above, we performed two separate gene-based collapsing analyses. The first collapsing analysis focused on LoF variants (frameshift, stop–gained, splice site donor and splice site acceptor). The second included LoF variants as well as non-frameshift indels and missense variants that are classified as 'probably damaging' according to PolyPhen HumDiv (≥0.957). Both gene-based analyses were performed following a dominant model, where a single damaging variant is assumed to be sufficient to lead to a deleterious effect on phenotype. We compared the proportion of cases that have one or more damaging variants to the proportion of controls with one or more damaging variants using a two-sided Fisher's exact test. We set the multiple test correction as 0.05 / (18,816 genes × 2 sets of tests) = $1.3 \times 10^{-6}$.

To determine for each of the single test sets above if there is evidence for genomic inflation, we used permutation-based assessment through the QQperm R package (https://cran.r-project.org/web/packages/QQperm/index.html). The QQperm package takes a gene × sample collapsing analysis matrix and derives expected chi-square statistics via permuting case–control labels and for each permutation, deriving the median chi-square statistic.

In the largest single case–control comparison (Caucasian ancestry, $n = 925$ cases and 4,340 controls), we did not observe a single gene-based result that passed the preset significance threshold. To make better use of the full data, we meta-analyzed collapsing results across all six available case–control groups, using the same $P$ value threshold as previously defined.

**Collapsing meta-analyses.** We combined collapsing analysis results across the 11 case–control groups to obtain single-gene $P$ values across the full data. To do this, for each single gene represented within the collection of analyses, we conducted a two-sided Cochran–Mantel–Haenszel test under the null hypothesis that, for a particular gene, the six underlying single analysis contingency tables form a corresponding weighted OR of 1 for the binary outcome. The Cochran–Mantel–Haenszel test is known to be robust to differing sample sizes across strata. In particular, the 'exact test' variant of this test that we conducted in R (the function 'mantelhaen.test(exact=TRUE)') should produce exact $P$ values across $2 \times 2 \times k$ tables, regardless of differences in case–control group sizes or strata sizes[45]. It has been used on exome sequencing case–control summary statistics in several peer-reviewed publications from our group[44,46] and has also been used on case–control statistics from studies of rare copy number variation from other groups[47,48]. To test for evidence of heterogeneity in the results, we used a Woolf test on the set of 11 contingency tables and considered any results with Woolf test $P < 0.05$ as having a degree of heterogeneity across the results that were suspect.

To generate QQ plots and genomic inflation factors for these results, we used an extended version of the QQperm approach. Over a total of 10,000 permutations, we 1) permuted case–control labels in each separate case–control group, 2) generated gene-based test statistics for the permuted dataset using the Cochran–Mantel–Haenszel test as described above and 3) derived the median chi-square statistic from the set of tests. All permutations and genomic inflation factor calculations were, again, performed using QQperm.

**Selection of a single-ancestry group of samples for case–control comparisons of LoF SNV/indel rate.** To make an accurate determination of the contribution of LoF SNVs/indels to OCD, we produced a single-ancestry case–control group that was more representative of a comparison of OCD to the general population. From the uniform manifold approximation and projection plot of all samples included in the gene-based collapsing analysis in Supplementary Fig. 1, it is apparent that a set of three neighboring clusters form the core of samples that are of self-declared European ancestry. These clusters were labeled in the data as cluster 0, cluster 3 and cluster 4. These three clusters also harbor three-fourths of the total cases in the full data. We took individuals from any of these three clusters (943 cases and 4,730 controls) and retained the subset that carried a phenotype listing of 'obsessive compulsive disorder', 'healthy family member' or 'control'. To remove outliers potentially remaining in the genetic data, we used EIGENSOFT[49] version 6.1.4 to generate the first ten PCs across the data and conduct iterative outlier pruning for a total of ten iterations. We were left with a total of 845 cases and 1,761 healthy

controls suitable for LoF rate comparison. A PCA of these samples suggests that ancestry in cases and controls are similar and that any remaining differences can be effectively controlled for by using critical PCs as covariates (Supplementary Fig. 4).

**Case–control comparisons of LoF SNV/indel rate.** Rate-based analyses were focused on variants that met the same quality and minor allele frequency (MAF) thresholds as described in gene-based collapsing. We retained variants for analysis if their annotation fell into one of the following two bins: 1) presumably neutral synonymous variation and 2) LoF variation with stop–gain, splice–donor, splice–acceptor or frameshift annotation. Because a centerpiece of these analyses is LoF variation in genes that are depleted of LoF variation in the general population, it was important that LoF variation in these genes be most likely real based on internal and external data and that they be near-absent from the general population. For this, we first removed variants with instances of low-quality calls across samples via the following additional constraints: ATAV filter classification of PASS (VQSR tranche < 90% for SNVs and VQSR tranche < 95% for indels); no QC status of 'fail' in any other case or control in the cohort (based on a call that failed VQSR criteria used in collapsing analyses); and gnomAD exome and genome FILTER classification of PASS. We then retained only variants if, in addition to earlier described MAF criteria, they could specifically be classified as ultra-rare, here defined as MAF < 0.00001 in the full gnomAD exome cohort and MAF < 0.00001 in the full ExAC exome cohort, equivalent to being found in no more than approximately three samples across the combined external data. To control for coverage differences, here we kept only the subset of qualifying variant calls where the binomial test *P* value for case–control genotype missingness based on insufficient coverage was greater than or equal to 0.05.

Basic comparisons of synonymous variant rate suggest that the datasets are similar. In comparing the mean synonymous variant count in cases versus controls, we found that a difference between the two was present but was marginal (average 12.3 in cases and 12.7 in controls, linear regression $P = 0.08$). Density plots of the counts across cases and across controls suggested that the small marginal difference stems from a small number of samples with a somewhat higher count that are preferentially found in the control group.

We set up regression models to control for covariates that could potentially confound tests of association between variant count and OCD case status. Before testing, we assembled a group of covariates that had potential to bias any case–control comparison made and tested each one for nominal evidence of association with the total exome synonymous qualifying variant count or with case phenotype ($P < 0.01$ for at least one of two associations). These included sex, the percent of CCDS release 20 bases covered at least 10×, the first 20 PCs and an indicator for instances of shared kits, where the exome kit used was represented in both case and control groups. We kept sex, percent of CCDS bases covered 10× and PCs 1 and 6 for our analysis. When the above covariates were included, the difference in synonymous variant count in cases relative to controls was negligible (covariate-adjusted difference = 0.09 and linear regression $P = 0.44$).

In tests specific to LoF variation in specific sets of genes, we retained the model described and extended it with synonymous variant count covariates. For tests of LoF variation within a specific set of genes, we formed regression models that included previously selected covariates (sex, percent of bases covered 10× and PCs 1 and 6) along with exome-wide synonymous variant count and synonymous variant count within the gene set. The critical predictor was the number of LoF variants within the test gene set per sample in the logistic regression model, whereas, in a linear regression model, this is the outcome, and OCD case status is the critical predictor. For each set of genes tested, we also assembled a simplified complementary test that compared the count of LoF variants relative to the count of synonymous variants, with the null hypothesis tested via a two-sided Fisher's exact test being that the proportions of LoF to synonymous variants will be equal in cases versus controls.

**Trio-based quality control.** Each trio was included only if each member passed individual level QC and displayed pairwise relatedness patterns that are expected for a trio family structure. Each individual sample was expected to meet the following QC thresholds: >85% of CCDS bases covered at least 10×; mean CCDS coverage > 10×; percent read duplication <60%; percent of reads aligning > 90%; and percent PCR contamination < 8%. For a trio to be included in an analysis, all three family members had to meet these criteria. In addition, as before, we used genotypes from 29,961 common variants that are well covered by the Roche version 3 kit to assess relatedness between samples. We used the kinship function of the KING version 1.4 statistical genetics software package to determine if the pairwise relationship patterns in each family are consistent with our data. For this, we required all parent–child pairs to have a KING relatedness coefficient between 0.17 and 0.35 and all parent pairs to have a relatedness coefficient < 0.10. All families included in this study passed the relatedness test described. A total of 587 trios and 41 quartets survived these QC procedures and were deemed suitable for DNM calling.

**Joint coverage assessment.** Trios that were a part of the compiled analysis cohort were assessed for their joint coverage statistics. Specifically, for each trio, we wanted to identify the subset of CCDS release 20 plus splice site bases that were covered ≥ 10× in each trio member, because this was a requirement that had to be met for a DNM call to be made. The subset of these sites across the exome for each sample are important to identify because they represent the true subset of the exome where a de novo SNV or indel call can be made.

We determined which genomic bases carried coverage ≥ 10× using mosdepth version 0.2.4, using the command 'mosdepth –threads 2 –by addjCCDSr20.bed –quantize 10:20:50:–no-per-base –fast-mode OUTROOT input.bam'. We then used bedtools version 2.25.0 to take the trio-level intersect between these loci across family members as jointly covered loci for that trio.

**DNM calling and filtering.** We pulled down a set of candidate DNM calls using ATAV 'list-trio' function. For the initial set of candidate DNM calls, the following QC thresholds were applied: coverage ≥ 10×; ≥ 3 reads supporting call; ATAV filter classification of PASS, LIKELY OR INTERMEDIATE (VQSR tranche < 99.9% for SNVs and VQSR tranche < 99% for indels); indel QD ≥ 2; indel FS < 200; indel ReadPosRankSum > −20; QUAL ≥ 30; QD ≥ 1; GQ ≥ 20; no status as an artifact based on internal records or EVS 'FAIL' assignment; coding region annotation in at least one CCDS transcript; and MAF < 0.01 in internal controls as well as EVS and ExAC global and subpopulation cohorts. To extract a set of high-confidence DNM calls from this lower-confidence callset, we, furthermore, required that calls meet the following criteria: variant classification of 'DE NOVO' via the ATAV 'list-trio' function; SNV MAF < 0.0005 in global and ancestry subpopulations of internal controls, EVS and gnomAD; indels absent from all internal and external controls (including EVS and gnomAD); no internal controls that have a QC fail for the variant call based on QC parameters for initial list-trio function call; status of 'pass' for variant call in proband; at least 30% of reads in child-supporting variant call; ≥ 10× coverage in proband and each parent; and MQ ≥ 40, QD ≥ 2 and QUAL ≥ 30 in proband. We additionally filtered out any call found in more than 5% of parent reads at the call site. Finally, we required that each included proband have no more than five DNM calls that met these criteria. Any probands that had a mutation count that exceeded this threshold were excluded from DNM analysis.

All DNMs were visually inspected in Integrated Genomics Viewer to make sure that the underlying alignment was sound. For each alignment, we visualized the span of bases 60 nucleotides upstream to 60 nucleotides downstream of the DNM call in each trio family member (father, mother and DNM carrier child). Variants that, based on visual assessment, are derived from a region with poor underlying read alignment were rejected.

**Sanger validation and estimated call accuracy.** To determine the accuracy of our callset, we selected a subset of SNVs and indels for Sanger validation. As criteria for inclusion in the validation set, a variant had to meet one of the following criteria: 1) LoF (stop–gain, splice donor–acceptor or frameshift), 2) part of recurrent non-synonymous gene-based signal or 3) missense with PolyPhen HumDiv ≥ 0.957 in a gene with RVIS < 25th percentile. A set of 106 de novo SNV and 28 indel calls were selected for Sanger validation.

We found that the validation rates for SNVs and indels were both high. Of our 106 SNV calls and 28 indel calls sent for validation, we were able to obtain data for 84 and 24 calls, respectively. Of these, 81 SNVs and 23 indels were assessable. We were able to validate 80 (98.8%) and 22 (95.7%) of these, respectively (Supplementary Tables 12 and 13). These percentages suggest that the large majority in our dataset that have not been assayed are likely real.

**Joint coverage in trio cohort data.** For each individual trio, we first identified all CCDS release 20 coding and splice donor–acceptor bases that were jointly covered at least 10× in every trio member. Based on the call criteria previously described, these loci would contain all bases where a DNM call could be made.

In general, joint coverage statistics suggest that we should have good power to call coding de novo SNVs and indels wherever they occur in the exome. We found that, across trios included, on average 96.1% of CCDS bases are jointly covered ≥ 10× (Supplementary Table 10). Furthermore, there were no major outliers in terms of poor coverage: only three trios of 587 had below 90% of bases covered, with a minimum of 87.6% of bases covered in a sample. Of the 18,852 genes included in CCDS release 20 loci, we found that, on average across trios, 95.1% of bases were jointly covered and that 16,230 genes (86.1% of total CCDS release 20 genes) had a mean percent of bases jointly covered ≥ 10× greater than or equal to 90% (Supplementary Table 11).

For each individual CCDS release 20 coding and splice site base, we summed up the total number of trios jointly covered at least 10×. We used these to compute mutation rate estimates for each gene, taking into account the number of trios covered at each base and broken down by possible coding annotation. For all rate-based tests focused on our 587 OCD trios, we should be optimally powered, because we are able to take into account the number of trios adequately covered to detect a DNM at each single base in producing mutation rate estimates.

**Mutation rate modeling.** To accurately model expected DNM rate, it was necessary for us to first produce a map of all possible coding annotations across the exome. Here we generated a VCF file containing all possible SNVs mapping to the human reference genome build 37 (within CCDS release 20 and splice site bases specifically) and annotated using clinEff version 1.0c and Ensembl 87 annotations.

We only considered annotations on the subset of Ensembl 87 transcripts tagged as 'CCDS' ($n = 37{,}309$). We first produced a table of expected mutation rate for each possible SNV, for easier gene-based and per-annotation summation, using a pre-computed DNA tri-mer mutation rate matrix (kindly provided by Shamil Sunyaev and Paz Polak). We next produced a table of maximum deleteriousness per SNV, with LoF SNVs being classified as '2', missense set classified as '1' and synonymous variants classified as '3'. Finally, we assigned PolyPhen-2 (ref. [11]) HumDiv scores to CCDS transcripts with SnpSift version 4.3i[50] using annotations from dbNSFP2.9 (refs. [51,52]) and recoded missense variants in this table with the maximum PolyPhen score, ranging from 0 to 1. Missense variants classified in this paper as 'damaging' (that is, 'misD') have a maximum PolyPhen-2 HumDiv score $\geq 0.957$, the cutoff for the classification of 'probably damaging' according to the dbNSFP manual, provided to them by the authors of PolyPhen-2.

Using the mutation rate and annotation matrices that we produced as described above, as well as the number of trios for each CCDS base jointly covered $\geq 10\times$, we went on to produce a rate estimate for each combination of coding gene and unique annotation (synonymous, missense and LoF). For a particular SNV at base $i$ with a clearly identifiable change in 'trimer' base content, the expected rate of DNM is set as $\lambda_i = m \times \Sigma(a \times t)$, where $a$ is a vector of allele copies per trio proband (always 2 if autosomal), $t$ is a vector where $t_j$ is set to 1 if the trio in question has joint coverage $\geq 10\times$ at the site and 0 otherwise and $m$ is the empirically derived mutation rate for the particular SNV you are looking at given the sequence trimer content. An expected mutation rate for any combination of gene and annotation class is computed as the sum of all $\lambda_i$ that map to the gene in question and carry that class as the maximum annotation in CCDS transcripts.

We also estimated frameshift rates per gene as a function of nonsense SNV rate. The expected rate of frameshift indels per gene was obtained by multiplying the expected nonsense mutation rate by 1.25, the ratio of observed singleton frameshift mutations to singleton nonsense mutations across 2,000 autism exomes, an adjustment introduced in ref. [19].

**Mutation rate tests.** For comparison of observed mutation counts in our cohort versus computed expected rates, we performed one-sided Poisson tests. We cross-checked our results with those on the same data from denovolyzeR[53], which uses the same tests but on different pre-computed rates per combination of gene and annotation group. We focused all exome-wide or gene set rate analyses on autosomes, because there will always be two underlying allele copies underlying every individual call site regardless of sample sex.

**Case–control burden.** For case–control comparisons, we used DNM calls from a callset derived from $n = 1{,}911$ unaffected siblings used in DNM analyses relative to their autism proband siblings[10]. We used the callset from Supplementary Table 2 of ref. [10] (specifically, variants listed as found within 'sF' or 'sM' only, which were healthy siblings of ASD probands), annotating the variants using the same pipeline as OCD probands. We required that variants in controls meet the same MAF criteria as those in cases (MAF < 0.0005 in gnomAD exome global and subpopulations). We also removed indels in both cases and controls found in an individual sample closer than 200 nucleotides in proximity, because, given the low frequency of de novo indels across samples, de novo indel calls that are this close to one another in a single sample are far more likely to be a result of poor read alignment. We controlled for potential differences in coverage by deriving a BED file consisting of CCDSr20 loci where > 90% of OCD probands are covered at least 10× and where > 90% of 709 internal controls sequenced on the same exome kit as the control trio samples used were also covered at least 10×. These loci should represent consensus regions where a DNM would be callable in either cases or controls. Cases were far more likely to have a DNM call outside of these loci ($P = 3 \times 10^{-8}$), likely due to sequencing being done on kits with more comprehensive coverage of the exome.

To effectively compare DNM burden in cases to that of controls using the information collected above, we used simple regression modeling of counts per DNM annotation. To estimate an OR for every additional DNM counted in a particular sample, we used a logistic regression model where phenotype was the outcome and DNM count per sample was the predictor. To estimate the difference in case rate versus control rate, we used a linear model where the sample-level DNM count was the outcome and phenotype was the predictor. For both models, to control for inferred differences in coverage, we defined the count of DNMs outside of jointly covered loci per sample as a covariate. We focused our analyses on autosomal genes to eliminate the possibility of sex differences influencing the number of bases where a mutation call could be made. Our well-controlled results in comparing case–control burden of presumably neutral synonymous variation suggests that this design is successful in controlling for potential coverage and sex differences in the data and is valid for comparisons of non-synonymous DNM burden (Fig. 4).

**Rate-based tests of LoF/misD burden in gene sets.** Using our sequence-based mutation rate model, we tested the burden of LoF and misD DNMs within three gene sets. The first consists of 187 genes and is derived from a union of (1) genes listed as ASD risk genes with $q < 0.1$ from ref. [16] ($n = 102$ genes) and (2) neurodevelopmental disease risk genes from ref. [20] with $P < 5 \times 10^{-7}$ ($n = 124$ genes). The second is a set of six Tourette's syndrome genes from ref. [8] with TADA $q < 0.3$ (classified by the author as 'probable risk genes') and impacted with at least two LoF/misD DNMs in the trio cohort. The third and final gene set consists of 199 genes from ref. [8] with only one single LoF/misD DNM hit.

**Case–control burden of misD DNMs in constrained regions.** We used MPC scores[12] to perform tests focused on the burden of presumably the most deleterious fraction of misD de novo SNVs. Here, we used qualifying missense variation with a pre-computed MPC score of at least 2. For comparisons focused on misD DNMs per MPC bin (0–1, 1–2 and >2), we considered only misD DNMs that were absent from the non-neurological subset of gnomAD version 2.1, an approach that, in a cohort of schizophrenia trios, was noted to enrich the missense DNM burden signal relative to expectation[23].

**Gene-based tests of observed versus expected DNM rate.** We used one-sided Poisson tests to compare the observed versus expected DNM rate within each gene. We did this for two sets of annotations: LoF-only and LoF/misD mutations. We used DNM calls across a total of 771 trios, including the 587 trios described within this analysis and an additional 184 OCD trios described in ref. [7]. The mutations from the additional 184 OCD trios were obtained from Supplementary Table 2 of ref. [7], specifically from samples where 'batch' was listed as 'ocd' and the 'exclude' column value was '0'. Mutation rates for LoF and misD annotations per gene were produced using the gene-based model described before, omitting base-level coverage information and assuming that all trios are adequately covered at each coding base to make a call. This was done because we had only the published DNM callset for the 184 trios from ref. [7], and determining the number of trios adequately covered per base would require raw sequence data. The expected mutation rate for each gene was the per-allele mutation rate multiplied by the total number of alleles present in the trio cohort (assumed to be $771 \times 2$).

**Joint analysis of DNM and case–control data using extTADA.** We used extTADA to extend the power of rate-based DNM tests with LoF variation from an independent case–control cohort. We used the cohort of singleton OCD cases and healthy controls, which we described producing earlier (476 cases and 1,761 controls). Selection of LoF variants for inclusion was done as it was in LoF rate tests described previously.

We ran extTADA following the code outlined at https://github.com/hoangtn/extTADA/blob/master/examples/extTADA_MultipleSteps.pdf and adding in calls to the case–control counts that were generated. Parameter estimation was done using the extTADAmcmc() function with the number of iterations set to 20,000. Parameter estimation led to the following average relative risk ($\lambda$) estimates (lower–upper credible intervals) for the following categories: misD DNM = 7.82 (1.05–30.20); LoF DNM = 20.97 (1.34–78.54); and LoF case–control = 2.83 (1.01–15.26). We also obtained estimates for variability in relative risk estimates per gene ($\beta$) and for the following: misD DNM = 0.88; LoF DNM = 0.82; and LoF case–control = 1.27. These derived parameters were used as input into the extTADA function calculateFDR() to obtain Bayes factors, posterior probabilities and $q$ values per gene. We define genes that have a $q$ value less than 0.3 (that is, significant with an FDR < 0.3) as probable risk genes and those that are less than 0.1 (that is, significant with an FDR < 0.1) as high-confidence risk genes.

**Statistical analysis.** With the exception of the extTADA analysis, all statistical analysis of genetic data was carried out in R version 3.2.3. The extTADA analysis software is primarily written in R, but, owing to compatibility issues centered on its dependencies, we ran it using R version 3.4.3. For almost all analyses, statistical tests were done using base R functions.

Gene-based collapsing meta-analysis of contingency tables was carried out using the function 'mantelhaen.test' with parameter 'exact=TRUE'. For each single gene, we provided as input to this function a three-dimensional stack of contingency tables across the six input groups, formed via the 'abind' R function from the 'abind' R package with parameter 'along=3'. We used the 'estlambda2' function from the QQperm R package to compute a genomic inflation factor (lambda) for the meta-analysis using case–control permutation on each input group, performing a total of 1,000 permutations. To test for evidence of heterogeneity in the results, we used the 'WoolfTest' function from the 'DescTools' R package.

Case–control comparisons of LoF variant and DNM rate were carried out using basic regression models in R. We used a logistic regression model to test for an association between variant count and the case–control status outcome, with covariates described in the main text included. We used a linear regression model to test for an association between case–control status and variant count as an outcome, with the same covariates included. Logistic and linear regression models were formed using standard R functions 'glm(family=binomial)' and 'lm()', respectively.

Tests of DNM rate versus expectation based on sequence content and the number of input trios were carried out using standard Poisson tests. We performed one-sided Poisson exact tests in R via the base function 'poisson.test' with parameter 'alternative=greater' to set up a one-sided test of the null being that the observed rate is less than or equal to expected.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Gene-based collapsing analysis summary statistics are provided in Supplementary Tables 4 and 5. DNMs detected across the 587 OCD trios and 41 quartets are provided in Supplementary Tables 8 and 9, respectively. Summary statistics from extTADA analysis (including LoF counts per gene in 476 OCD cases versus 1,761 healthy controls) are provided in Supplementary Table 16. Clinical data for cases and healthy family members sequenced as part of this study are available on dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000903.v1.p1).

## Code availability

Extraction of sample-level coverage information and extraction of QC-passing genotypes was done using ATAV (https://github.com/igm-team/atav). Manipulation of PLINK files and subsetting according to genotype missingness were done using PLINK version 1.90_3.38 (https://www.cog-genomics.org/plink2/). Kinship analysis was performed using KING version 1.4 (http://people.virginia.edu/~wc9c/KING/). PCA was performed using FLASHPCA version 2.0 (https://github.com/gabraham/flashpca) in gene-based collapsing analysis and EIGENSOFT version 6.1.4 (https://data.broadinstitute.org/alkesgroup/EIGENSOFT/) in LoF rate comparisons. Calculation of trio-based coverage was done using mosdepth version 0.2.4 (https://github.com/brentp/mosdepth) and bedtools version 2.25.0 (https://github.com/arq5x/bedtools2/releases). Full analysis code written specifically for the analysis described in this manuscript is available in the Supplementary Software Appendix and is also available in a public repository (https://github.com/Halvee/OCD_WES_analysis_full_NatureNeuro2020).

## References


38. Guze, S. B. *Diagnostic and Statistical Manual of Mental Disorders, 4th ed.* (DSM-IV). *Am. J. Psychiatry* **152**, 1228–1228 (1995).
39. Mannuzza, S., Fyer, A. J., Klein, D. F. & Endicott, J. Schedule for Affective Disorders and Schizophrenia–Lifetime Version modified for the study of anxiety disorders (SADS-LA): rationale and conceptual development. *J. Psychiatr. Res* **20**, 317–325 (1986).
40. Glasofer, D., Brown, A. J., & Riegel, M. *Structured Clinical Interview for DSM-IV (SCID).* (Springer, 2015).
41. Raghavan, N. S. et al. Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann. Clin. Transl. Neurol.* **5**, 832–842 (2018).
42. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
43. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
44. Povysil, G. et al. Assessing the role of rare genetic variation in patients with heart failure. *JAMA Cardiol.* **6**, 379–386 (2021).
45. Agresti, A. A survey of exact inference for contingency tables. *Stat. Sci.* **7**, 131–153 (1992).
46. Cirulli, E. T. et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436–1441 (2015).
47. Rees, E. et al. Analysis of intellectual disability copy number variants for association with schizophrenia. *Nat. Genet.* **49**, 1167–1173 (2017).
48. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
49. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
50. Ruden, D. et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012).
51. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
52. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
53. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15 (2015).


## Acknowledgements


We are grateful to the research participants, as this study would not have been possible without their participation. This work was supported by the following National Institutes of Health grants: 'Identifying de novo mutations causing OCD in trios by whole exome sequencing' (MH099216—D.B.G. and G.N.); 'Identification of rare variants of OCD' (MH097971—D.B.G.; MH097993—G.N.). Data acquisition was made possible by the OCD Collaborative Genetics Association Study, funded by the following National Institute of Mental Health grants: MH071507 (D.G.), MH079489 (A.J.F.), MH079487 (J.M.), MH079494 (J.A.K.) and MH 071507 (G.N.).


## Author contributions

D.B.G. and G.N. conceived of and obtained funding for the research, and D.B.G. designed the experiments. J.F.S., J.K., D.G., A.J.F., B.D.G., J.T.M., O.J.B., J.A.K., M.A.R., M.A.G., P.S.N., Y.W., F.S.G. and B.M. collected samples, prepared samples for analysis or were involved in clinical evaluation. IGM staff members performed all experiments, and M.H. executed data analyses, with critical help provided by T.D.P. D.B.G., G.N., B.M., T.D.P., A.W.Z. and F.S.G. provided analysis suggestions. M.H. and D.B.G. performed the primary writing of the manuscript, with input from G.N., B.M., F.S.G., A.W.Z. and J.F.S. All authors approved the final manuscript.

## Competing interests

D.B.G. reports equity holdings in precision medicine companies and consultancy payments from Gilead Sciences, AstraZeneca and GoldFinch Bio. All other authors declare no competing financial interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00876-8.

**Correspondence and requests for materials** should be addressed to G.N. or D.B.G.

**Peer review information** *Nature Neuroscience* thanks Ditte Demontis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

nature research

Corresponding author(s): David B. Goldstein
Gerald Nestadt

Last updated by author(s): May 2, 2021

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Aligned reads were generated using the Dynamic Read Analysis for Genomics (DRAGEN) platform, as previously described (PMID 30009200). Variant calls were generated from aligned read data using GATK v3.6 (https://github.com/broadinstitute/gatk). Variant calls that met pre-set QC thresholds across case/control and trio-based contexts were extracted using ATAV (https://github.com/igm-team/atav). Variant annotation was done by ATAV using clinEff v1.0c on Ensembl 87 transcripts. Manipulation of PLINK files and subsetting according to genotype missingness was done using PLINK v1.90_3.38 (https://www.cog-genomics.org/plink2/). Pairwise relatedness between samples was calculated using KING software package v1.4. PCAs were performed using FLASHPCA v2.0 (https://github.com/gabraham/flashpca) in gene-based collapsing analysis and EIGENSOFT v6.1.4 (https://data.broadinstitute.org/alkesgroup/EIGENSOFT/) in LoF rate comparisons. Scripts for cleaning and formatting the data before analysis were written in R v3.2.3 and Python v2.7.7. For calculation of coverage, coverage summary statistics in case/control comparisons were generated by ATAV. Trio-based coverage statistics were generated using a combination of mosdepth v0.2.4 (PMID 29096012) and bedtools v2.25.0 (PMID 20110278). |
| Data analysis | ATAV (PMID 33757430) was used to extract coverage-harmonized variant calls used in case/control analyses, as well as de novo variant calls present in trio probands and absent in parents. Code written specifically for the analysis of these variants described within this manuscript has been made available in the supplementary software appendix, as well as on github (https://github.com/Halvee/OCD_WES_analysis_full_NatureNeuro2020). All statistical analyses described were conducted in R v3.2.3, with the exception of extTADA analysis (PMID 29262854, https://github.com/hoangtn/extTADA), which was conducted in R v3.4.3. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Gene-based collapsing analysis summary statistics are shared in supplemental tables 4 and 5. De novo mutations detected across the 587 OCD trios and 41 quartets are provided in supplemental tables 8 and 9 respectively. Summary statistics from extTADA analysis (including LoF counts per gene in 476 OCD cases versus 1761 healthy controls) are provided in supplemental table 16.
Minor allele frequencies from the following external datasets were utilized in this study:
gnomAD v2.1 : https://gnomad.broadinstitute.org/
ExAC v0.3 : https://gnomad.broadinstitute.org/
EVS v0.0.30 : https://evs.gs.washington.edu/EVS/

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | Sample sizes for all experiments done in this study were the maximum that were attainable given funding, recruitment efforts and access to additional control sample data at the time of analysis. For gene-based collapsing analysis the total number of samples were determined by collecting all sequenced OCD cases and applicable non-neuropsychiatric controls available at time of analysis, subsetting on those that met stringent QC criteria, forming sample clusters from genetic data reflective of ancestry (11 total), and then counting the number of cases and controls per group that remain and were included in comparisons. For case/control LoF rate comparison, we took 3 neighboring clusters that mapped to European ancestry in our gene based collapsing analysis, and subsetted on OCD cases and explicitly healthy controls found within these clusters. For de novo mutation analysis, the total number of case trios and quartets were determined by collecting all sequenced OCD trios and quartets available at time of analysis that met stringent QC criteria. The total number of control trios were taken from the reported total number of healthy siblings of ASD probands as reported in Iossifov et al. (PMID 25363768). |
| Data exclusions | Samples were excluded from both case/control and trio analyses if there was evidence for overall poor sequence data quality based on overall coverage statistics, read duplication, percent of reads aligning to the genome and PCR contamination metrics. Samples were only included in case/control analyses if they did not have sufficient evidence for cryptic relatedness with any other cases or controls in the cohort. Full trios were only retained in trio analyses if they were unrelated to other trio members, and if within the trio, relatedness patterns were consistent with the family structure (mother/father non-relatedness, mother/child relatedness, father/child relatedness). |
| Replication | For case/control studies, there was not a comparably sized independent cohort of OCD exome sequence data available at time of analysis for replication of findings specific to our analysis. For case/control comparisons of LoF rate, we observed an excess of LoF rate in LoF-intolerant genes, which is consistent with prior findings across a spectrum of separate neuropsychiatric disorders. We observed an excess of damaging de novo SNVs and indels in the described cohort of 587 trios relative to controls, and found that this result was also observed in a comparison against expected mutation rate. In general, our detection of an excess of damaging de novo SNVs and indels in this cohort is a replication of similar findings from a previously published smaller cohort of 184 OCD trios (PMID 31771860). We also used Sanger sequencing to independently validate de novo mutation calls made from exome sequence data in OCD trios. |
| Randomization | All between-sample comparisons in our study involved allocation of samples into either a case group (defined as those with confirmed OCD) or a control group (defined as those without confirmed OCD). In gene-based collapsing analyses, 11 separate case/control groups were defined based on clustering reflecting of ancestry. In LoF rate comparisons, only samples from 3 neighboring clusters forming the core of European ancestry individuals were used, and covariates pertaining to coverage and population structure were included in regression models to control for subtle influences on overall rare coding variant rate. In DNM rate comparisons, as before all cases were grouped together and all external controls were grouped together. The number of mutation calls per sample falling outside of joint capture loci were used as a covariate to account for differences in coverage between groups attributable to differences in capture kits used. |
| Blinding | During the recruitment process, investigators were focused entirely on the collection of samples from OCD cases and their families, and were not blinded to phenotype. Since this study consisted of both cases and controls, and in analysis it was important to know which group each sample belonged to in order to control for confounders, investigators were not blinded to group identity during analysis. Control samples were selected to match cases based on sequencing quality and ancestry. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | Cases recruited in this study all have a primary diagnosis of OCD. Of the 1313 OCD cases included in analyses, 554 are male and 759 are female. Cases are predominantly of self-declared European ancestry (1177/1313). Among the 1313 cases, the mean age at recruitment was 30.6.<br><br>We have also produced population characteristics across each of the three case analysis groupings (collapsing analysis, LoF rate, de novo mutation analysis trios). The collapsing analysis cases (n=1263) consist of 536 male and 727 female individuals, with a mean age at recruitment of 30.4. The LoF rate analysis cases (n=845) consist of 329 male and 516 female cases, with a mean age at recruitment of 32.0. The trio probands that were a part of the de novo mutation analysis (n=587) consist of 286 male and 301 female individuals, with a mean age at recruitment of 23.7. |
| Recruitment | All the samples were ascertained as part of the OCD Genetics Association Study (OCGAS) or at the Johns Hopkins University OCD clinic. The trios and quartets included in the study consisted of unaffected parents and where possible families with no additionally known affected relatives. The families were screened for evidence of OCD. All affected cases were examined by a PhD research psychologist using the Structured Clinical Interview for Diagnosis (SCID). Where possible an informant was also interviewed. All cases were reviewed independently by two research diagnosticians to establish diagnostic concordance. The participants in this study were recruited opportunistically from several sources (outlined in the online methods). Consequently, the findings may not reflect the population-wide characteristics of OCD cases. Potential ascertainment biases may be present, but these are not apparent given the demographic similarities of this sample to those identified in population-based studies. |
| Ethics oversight | The study was reviewed and approved by the Institutional Review Boards at the five sites where participants were recruited (Butler Hospital, Columbia University, Johns Hopkins University, Massachusetts General Hospital, and the University of California at Los Angeles). |

Note that full information on the approval of the study protocol must also be provided in the manuscript.